Attorney Docket No.: 2558B-063700US

PATENT APPLICATION

PATTERN RECOGNITION METHOD FOR DIAGNOSIS OF SYSTEMIC AUTOIMMUNE DISEASES

Inventors: STEVEN R. BINDER, a citizen of the United States of America, residing at:
2506 Hawthorne Terrace
Berkeley, California 94708

JOHN GLOSSENGER, a citizen of the United States of America, residing at:
105 Rankin Way, No. 87
Benicia, California 94510

Assignee: Bio-Rad Laboratories, Inc.
1000 Alfred Nobel Drive
Hercules , California 94547
a corporation of the state of Delaware

Prepared by: M. Henry Heines
TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 415-576-0200

PATTERN RECOGNITION METHOD FOR DIAGNOSIS OF SYSTEMIC AUTOIMMUNE DISEASES

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention arises in the field of autoimmune diseases and diagnostic methods for these diseases. The invention also relates to statistical methods of data analysis and their application to immunodiagnostics.

2. Description of the Prior Art

Autoimmune diseases are conditions in which the immune system attacks cells, tissues, and organs of one's own body rather than bacteria, viruses, and other microbes that invade the body from the outside. There are many different autoimmune diseases, attacking different parts of the body ranging from the gut to the brain, for example, or from the vascular system to the skin, and the attacks occur in different ways. Some autoimmune diseases are tissue- or organ-specific, while others affect several tissues. Diseases of this latter category are termed "systemic autoimmune diseases," and the symptoms may vary from one patient to the next, with tissue injury and inflammation occurring in multiple sites in organs without relation to their antigenic makeup. Examples of systemic autoimmune diseases are rheumatoid arthritis, systemic lupus erythmatosus, scleroderma, polymyositis, dermatomyositis, Sjögren's syndrome, and spondyloarthropathies such as ankylosing spondylitis. Autoimmune diseases are generally multifactorial in origin, with some of the contributing factors being genetic disposition, host factors (such as T cell defects and polyclonal stimulation of B cells that are resistant to controls), environmental factors (such as certain microbial infections), and

antigen-driven mechanisms (such as sequestered antigens or cross-reacting exogenous antigens).

Due to their overlapping symptoms and complex etiologies, autoimmune diseases are difficult to diagnose. Attempts at diagnosis are generally based on symptoms, together with the findings from a physical examination and results obtained from laboratory tests. Symptoms of many autoimmune diseases are nonspecific, and while laboratory test results may help they are often inadequate to confirm a diagnosis, particularly in the early phases of the disease. The problem is aggravated when the symptoms are transient and the laboratory results are inconclusive, as they are in many cases. For some autoimmune diseases, the patient will respond completely to treatment if the disease is identified at an early stage. Often, however, a specific diagnosis cannot be made until the disease is in an advanced state. The problem is particularly acute with systemic autoimmune diseases.

It is commonly believed that the presence of autoimmune antibody is indicative of an autoimmune disease. This belief is disproved however by the fact that almost all antibodies are present in measurable amounts in any individual, and comparisons with reference levels from healthy individuals are frustrated by the wide variations in normal antibody levels due to demographics such as gender, age, geographical region of domicile, nationality and race. In addition, the number of differentiable antibodies that must be investigated in the diagnosis is large, which presents an often unmanageable burden on the clinical laboratory in its attempts to obtain a specific and reasonably accurate diagnosis.

The most widely used method of identifying a systemic autoimmune disease is indirect immunofluorescence, a manual method that requires a well trained technician. The result appears as a pattern of distinctive characteristics that are seen on cells (typically HeLa cells) that have been fixed on a slide, treated with serum, washed and then labeled. The distinctive characteristics of the pattern are a relatively speckled, relatively homogeneous appearance, or the like, and diagnosis is achieved by comparing the pattern with those of individuals with known diseases. The most common patterns are associated with several diseases, however, which obscures the diagnosis. As a result, additional testing for specific antibodies is needed. Even then, an experienced immunologist or rheumatologist must be called upon to interpret the results and make the final diagnosis.

Recently, enzyme-linked immunosorbent assays (ELISAs) have become available for preliminary screening of patients with symptoms suggesting an autoimmune disease. Specimens reading positive in the screening assay are then submitted for a specific diagnostic assay to determine which particular autoantibody is present and therefore which specific disease is suggested. Many autoimmune disease patients, however, suffer from two or more different autoimmune diseases, and this confounds efforts at achieving an accurate diagnosis when using this diagnostic method.

SUMMARY OF THE INVENTION

The difficulties enumerated above and others are addressed by the present invention, which resides in the identification of a particular systemic autoimmune disease (or diseases) by obtaining multianalyte test data from a patient suspected of suffering from a systemic autoimmune disease, applying a statistical pattern recognition method to this data to compare it with a multitude of reference data sets, and using the comparison to identify the specific disease(s) that the patient is suffering from. With the aid of computer software, the pattern recognition method processes the entire pattern of results in a medical decision support system. Applying pattern recognition methods in this manner permits the clinician to use test results from a single biological sample obtained from the patient, and to obtain both a diagnosis of the disease(s) and an assessment of the confidence level of the diagnosis. The autoantibodies selected can be those that are known to be frequently elevated in various systemic autoimmune diseases, and particular antibodies can be included or omitted from the set in accordance with their perceived relevance to the particular disease or group of possible diseases. Two or more diseases can be diagnosed simultaneously, and confidence levels assigned to each, without the need for separate samples or analyses.

The method of this invention offers a number of advantages over diagnostic methods currently known. One advantage is the elimination of the need for a manual assay, since the multianalyte assay itself can be performed by automated instrumentation. Another is the ability to obtain a diagnosis without the need for screening followed by confirmatory tests. In addition, the invention provides a diagnosis among a large number of possible diseases with a single-step sample analysis. Further, the invention eliminates the need for the intervention of a skilled professional to obtain an interpretation of the result. The result can therefore be transmitted directly to the ordering

physician, even if that physician is a generalist and unfamiliar with the autoantibodies being measured.

These and other objects, advantages, features and embodiments of the invention will be apparent from the description that follows.

DETAILED DESCRIPTION OF THE INVENTION AND SPECIFIC EMBODIMENTS

Pattern recognition systems for use in the practice of this invention typically begin with the development of a "training set," a term understood in the art of statistical data analysis to mean data from a set of samples from reliable ("pedigreed") sources. The set will include samples having disease conditions that are known from a previous and independent diagnosis as well as samples that are disease-free. The method in accordance with this invention thus begins with reference samples from a series of subjects each of whom is known to have, or to have had, a particular systemic autoimmune disease, as well as a series of subjects each of whom is known to be disease-free. This training set includes the full scope of the systemic autoimmune diseases sought to be investigated for a particular patient or for patients in general. Multianalyte analyses are performed on each sample in the training set, and the results are entered into a database in a manner resembling a multi-dimensional array in which each sample can be thought of as a point in a multi-dimensional space. The coordinates of the point that define its location in the space represent the value of the test results, one coordinate for each test. As a simple case, if the number of tests performed per sample in the training set is only two, the training set points will be arranged in a two-dimensional (x, y) plot, where the horizontal (x) axis corresponds to one of the tests and the vertical (y) axis corresponds to the other. In addition to its location in the two-dimensional plot (i.e., the x and y values) which is determined by the test results, each point is labeled as to the particular disease that is associated with its source. The corresponding array for a three-test system is a three-dimensional plot with orthogonal x, y, and z axes. Corresponding arrays for systems involving four or more tests are not capable of visualization, but are established in an analogous manner.

Once the training set is established and its test data entered into the database, the system is ready to receive data from a single patient sample or from a succession of patient samples with no need to recreate the database, or to re-analyze the

training set, or to select and analyze a different training set. The patient sample is subjected to the same tests as the samples in the training set, and these test results are inserted into the database. Using the illustrative representational description of the preceding paragraph, the patient sample test results are added to the space as a point at a location defined by coordinates equal to the test values.

The determination of particular diseases is then achieved by a statistical comparison between the values of the patient sample test results and those of the training set. Various methods of statistical analysis can be used for the comparison. Examples are k-nearest neighbor analysis, multi-linear regression analysis, Bayesian probabilistic reasoning, neural network analysis, and principal component analysis. While each of these methods is known in the art, the following explanations will assist the reader in understanding how the algorithm of each is applied in the practice of the present invention.

The k-nearest neighbor algorithm reads the data from the patient sample into memory which also contains the database of the training set. The algorithm then calculates the numeric distances from the patient's test point to the data points in the memory from the training set that are in the vicinity of the patient's test point, and from those points selects the *k* whose distances are the shortest. Thus, if *k* is 15, the algorithm selects the 15 that are the shortest distances from the patient's test point. The disease associated with the *k* nearest data points is then identified as that which is present in the patient sample, and if the *k* points are associated with more than one disease, the diagnosis is for each of the diseases indicated. A refinement for the algorithm is one in which the distance values of the various points are compared, and when the values for points representing two different diseases differ by less than a minimum difference, both diseases are considered to be equally likely. In a further refinement, the algorithm processes the numerical values of the distances to determine a confidence level. This can be done by using "similarity" values between the patient test results and each of the *k* data points and assigning a relative value to each disease by dividing the similarity for that disease by the sum of similarities. The final diagnosis is generally selected as that disease (or those diseases) with the highest relative value. In general, the k-nearest neighbor algorithm is a "case-oriented" system, or one that compares the patient's test data to the data from other individuals for whom demographic and other personal information is available in addition to the disease, such as age, sex, the length of time that the individual has had the disease, and similar factors. This information permits a diagnosing physician

to exercise individual judgment as to whether or not to use the patients proposed by the algorithm as reference points for the diagnosis.

A Bayesian algorithm differs from the k-nearest neighbor algorithm by calculating results in terms of probabilities based on all of the data in the training set, rather than a limited group selected on the basis of its proximity to the test patient. A Bayesian algorithm extracts two types of data from the data points in the training set database, one representing the disease prevalence and the other representing statistical measurements such as mean and standard deviation. A probability density is determined for the patient sample and each disease, and the likelihood of each disease is calculated. The relative value of each disease is then determined from a ratio of the likelihood for that disease divided by the total of the likelihoods for all of the diseases, and the diagnosis is achieved by selecting the disease with the highest relative value. Like the k-nearest neighbor algorithm, Bayesian algorithms can be used to detect the presence of two diseases in a single patient.

The neural network system is a network of nodes or simple numerical processing units that are arranged in layers and connected by communication channels. The channels are weighted differently in accordance with information imported from the training set. Data are passed from the first layer to successive layers in accordance with the different weights assigned to the channels. Each node in a given layer multiplies all the node values from the previous layer by the weights of the connection channels that join the nodes of the two layers, and then determines the sum of these values. The sum is then passed through a sigmoid ("S-curve") function to identify the disease.

Principal component analysis is a multivariate technique for reducing matrices of data to their lowest dimensionality by use of orthogonal factor space. According to this technique, the training set is processed to identify the number of principal components. This information is then used to model the patent test data using such techniques are target transformations or curve fitting. Neither the principal component nor the neural network analyses are suitable for the detection of two diseases.

Other pattern recognition techniques that are known to those skilled in statistical data analysis can be used as well, and their adaptation to the systemic autoimmune diseases addressed by this invention will be readily apparent to the skilled artisan. Regardless of the pattern recognition technique that is used, the technique and its application in this context are readily susceptible to computerized implementation.

Software specifically designed for particular analyses is readily developed and a routine matter to the skilled software engineer.

The test data used in the present invention are values that are proportional to, or otherwise representative of, the levels of various antibodies that are associated to various degrees with systemic autoimmune diseases. Currently, over 100 antibodies are known to be expressed in autoimmune diseases. Examples are listed in Peter, J.B., et al., *Autoantibodies*, Elsevier Science B.V., Amsterdam (1996), the contents of which are incorporated herein by reference. The antigens to many of these antibodies are commercially available, while the antigens to others are readily synthesized based on descriptions of them that are available in the literature. Some of the sources of these antigens are BiosPacific, of Emeryville, California, USA; Immunovision, of Springdale, Arkansas, USA; and KMI Diagnostics, Inc., of Minneapolis, Minnesota, USA. Examples of the antibodies that are expressed in autoimmune diseases, identified by the antigens to which they bind in an immunoassay, are listed below:

SSA 60
SSA 60
SSA 52
SSB 48
Sm BB'
Sm D1
RNP 68
RNP A
RNP C
Fibrillarin
Riboproteins P0, P1, and P2
dsDNA
Nucleosome
Ku
Centromere A
Centromere B
Scl-70
Pm-Scl
RNA-Polymerases 1, 2, and 3
Th

Jo-1
Mi-2
PL7
PL12
SRP

The present invention is not intended to be limited to antibodies identified by the antigens in the above list. Instead, the number of antibodies that are detected and quantitated from the training set, and the number of antibodies detected and quantitated in the patient sample as well, may vary and are not critical to this invention. The particular antibodies selected and the number of such antibodies in the selected group will however affect the accuracy of the assay and its ability to identify the systemic autoimmune disease(s) present. In most cases, it is contemplated that from 10 to 100 antibodies will be used, preferably from 15 to 25. In preferred embodiments of the invention, at least 15 of the antibodies identified by the antigens in the above list will be used, and in further preferred embodiments, all of the antibodies in the above list will be used.

Similarly, the number of reference samples of different sources used to develop the training set may vary as well and is not critical to this invention, although the number selected may affect the range of autoimmune diseases that can be detected. In most cases, it is contemplated that 100 to 10,000 reference samples will be used, preferably from 200 to 2000.

The quantitative levels of each antibody are determinable by conventional antibody assays. Immunoassays are generally preferred. Both antigen capture (indirect immunoassays) and class capture assays can be used, antigen capture being preferred, and detection can be achieved by enzyme labels, radioactive labels, or fluorescent labels. The assays may be colorimetric, luminometric, fluorometric, enzymetric, radiometric, or other methods such as nephelometry, turbidimetry, particle counting, or visual assessment. Examples of enzyme labels are alkaline phosphatase, β-D-galactosidase, glucose-6-phosphate dehydrogenase, horseradish peroxidase, β-lactamase, melittin, and urease. Examples of radioisotopic labels are cobalt-57 and iodine-125. Examples of fluorescent labels are succinimidyl esters and vinyl sulfones of xanthenes, cyanines, coumarins, benzimides, phenanthridines, ethidium dyes, acridine dyes, carbazole dyes, phenoxazine dyes, porphyrin dyes, quinoline dyes, and naturally occurring protein dyes. Fluoresceins and rhodamines are particular types of xanthene dyes. Specific examples are

6-carboxyfluorescein, 6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein, N,N,N',N'-tetramethyl-6-carboxyrhodamine, 6-carboxy-X-rhodamine, 5-carboxyrhodamine-6G, 5-carboxyrhodamine-6G, tetramethylrhodamine, Rhodamine Green, and Rhodamine Red. Umbelliferone is an example of a coumarin. Hoechst 33258 is an example of a benzimide dye. Texas Red is an example of a phenanthridine dye. Examples of cyanine succinimidyl ester dyes are sulfoindocyanine succinimidyl esters, (carboxyalkyl)cyanines succinimidyl esters, and BODIPY succinimidyl esters (Molecular Probes, Inc., Eugene, Oregon, USA). Examples of naturally occurring protein dyes are phycoerythrins. Many other examples will be readily apparent to those skilled in the art.

The sequence and manner of performing quantitative assays of multiple antibodies in the practice of this invention may vary widely. In preferred embodiments, the assays are performed simultaneously in a single sample by a multiplexed system that differentiates the assays from each other and thus provides individual values for each of the antibodies. This may be achieved in a variety of ways. Heterogeneous binding assays, in which one of the binding members is immobilized on a solid phase, are the most efficient, since they provide a ready means for separating the bound pairs from unbound species. Multiplexed heterogeneous binding assays in which solid particles, preferably magnetic microbeads, are used as the solid phase, are illustrative examples. In these assays, the beads are sorted into groups, each group containing the reagents for a single assay covalently bonded to the bead surface, the various groups differentiable from one another by virtue of distinguishing characteristics that permit separate detection of the assay result in one group from those in the other groups. The differentiating parameter may be particle size, fluorescence (independent of the fluorescent label used in the assay, when fluorescent assay labels are in fact used), light scatter, light emission, or absorbance. The binding member that is covalently bonded to the bead surface will vary depending on the type of immunoassay that is being used. Since the analytes in accordance with this invention are all antibodies, the antigens by which the antibodies are defined serve as particularly convenient binding members bonded to the bead surface. Actual differentiation of the bead groups is readily achieved by flow cytometry.

When magnetic beads are used, the magnetic character permits quick separation of the solid and liquid phases and convenient washing of the solid phase. A magnetic character can be imparted to the beads by using beads made of paramagnetic materials, ferromagnetic materials, ferrimagnetic materials, or metamagnetic materials. The magnetically responsive material is preferably only one component of the bead, the

remainder consisting of a polymeric material to which the magnetically responsive material is affixed or otherwise combined. The polymeric material can be chemically derivatized to permit attachment of the antigen or other assay reagent that enters into the binding reaction.

Multiplex systems of this description as well as others useful in the practice of this invention are disclosed in pending United States patent application no. 09/302,920, filed April 30, 1999, entitled "Multiplex Flow Assays Preferably With Magnetic Particles as Solid Phase," Michael I. Watkins et al., inventors. The entire disclosure of application no. 09/302,920 is incorporated herein by reference.

The biological samples on which the tests are performed can be any biological fluid extracted from the patient that is likely to have a detectable antibody population that is characteristic of the disease(s) sought to be investigated. Serum, plasma, urine, or cerebrospinal fluid may be used. Serum samples are preferred.

This invention is useful in the detection and diagnosis of systemic autoimmune diseases in general. Examples are systemic lupus erythmatosus (SLE), Sjögren's syndrome, scleroderma, polymyositis and dermatomyositis, CREST (*c*alcinosis, *R*aynaud's phenomenon, *e*sophageal involvement, *s*clerodactyly, and *t*elangiectasis), mixed connective tissue disease, Wegener's disease, rheumatoid arthritis, and spondylarthropathies.

The foregoing is offered primarily for purposes of illustration. Further modifications and variations of the various parameters of the composition and method of this invention will be readily apparent to those skilled in the art. For example, the pattern recognition algorithm can be adjusted to control the rate of false negative results, by setting a boundary that differentiates "healthy" from "unhealthy" for a particular disease at a higher level than the corresponding boundary for other diseases, thereby lowering the number of samples that will indicate the presence of the disease. These and other variations are included within the scope of the invention.